BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

CLIPPEDIMAGE= JP407044567A

PAT-NO: JP407044567A

DOCUMENT-IDENTIFIER: JP 07044567 A

TITLE: DOCUMENT RETRIEVAL DEVICE

PUBN-DATE: February 14, 1995

INVENTOR-INFORMATION:
NAME
SATO, OSAMU

ASSIGNEE-INFORMATION:
NAME COUNTRY
FUJITSU LTD N/A

APPL-NO: JP05188243

APPL-DATE: July 29, 1993

INT-CL (IPC): G06F017/30

ABSTRACT:

PURPOSE: To provide a document retrieval device capable of obtaining an absolutely sufficient retrieved result with the retrieval of one time by retrieving similar documents from a document data base with the document itself as a retrieval key.

CONSTITUTION: This document retrieval device is constituted of a retrieval key word set generation means 2 for analyzing an input document 1 and generating a retrieval key word set 3 for which weighing corresponding to document component elements is performed and a document retrieval means for retrieving the document data base based on the retrieval key word set 3, calculating the weight of respective matched key words for each document obtained as a result

and obtaining cumulative weight for the document of the retrieved result.
Since the cumulative weight indicating the degree of similarity with the input
document is added to the retrieved result, a user can efficiently select the
retrieved result by referring to it.

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号

特開平7－44567

(43)公開日　平成７年(1995)２月14日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　17/30				
		9194－5Ｌ	Ｇ０６Ｆ　15/ 403　３２０　Ｚ	

審査請求　未請求　請求項の数２　ＯＬ　（全 12 頁）

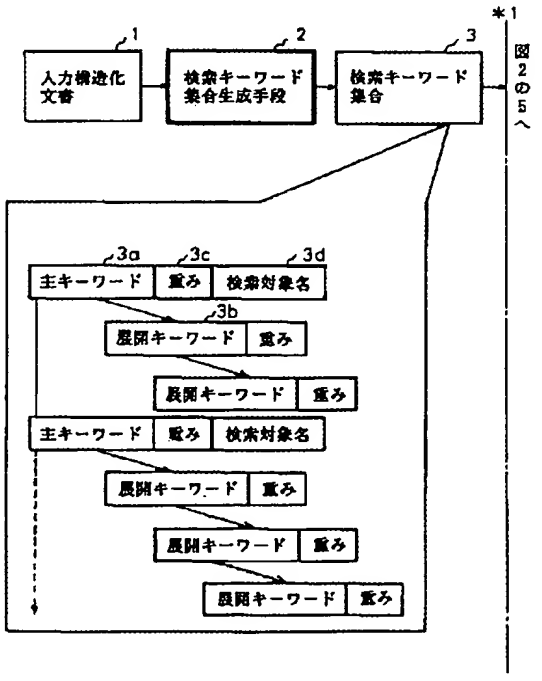
(21)出願番号　特願平5－188243

(22)出願日　平成５年(1993)７月29日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中1015番地
(72)発明者　佐藤　理
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内
(74)代理人　弁理士　宇井　正一　（外４名）

(54)【発明の名称】　文書検索装置

(57)【要約】
【目的】　文書データベースから、文書そのものを検索キーとして類似文書を検索し、一回の検索で必要十分な検索結果を得る文書検索装置を提供する。
【構成】　入力文書１を解析し、文書構成要素に従った重み付けをした検索キーワード集合３を生成する検索キーワード集合生成手段２と、前記検索キーワード集合３に基づき文書データベースを検索して、その結果得られた文書ごとに、マッチした各キーワードの重みを計算し、検索結果文書に対する累計重みを得る文書検索手段とから文書検索装置を構成する。検索結果には、入力文書との類似度を表す累計重みが付加されているので、利用者は、これを参考とすることにより、検索結果の取捨選択を効率的に行うことができる。



1

【特許請求の範囲】
【請求項1】　文書を格納した文書データベースから、利用者により入力された文書と類似の内容を持つ文書を検索する文書検索装置において、利用者が入力した定型的な構造を持つ入力構造化文書（1）を解析し、文書構成要素に従った重み付けをした検索キーワード集合（3）を生成する検索キーワード集合生成手段（2）と、
前記検索キーワード集合（3）に基づき文書データベース（4）を検索して、その結果得られた文書ごとに、マッチした各キーワードの重みから、検索結果文書に対する累計重みを得る文書検索手段（5）とを具備したことを特徴とする文書検索装置。
【請求項2】　前記文書データベース（4）に格納される文書を定型的な構造を持つ文書とし、前記検索キーワード集合生成手段（2）は、前記検索キーワードの重み付けを、入力構造化文書（1）の文書構成要素と、対応する前記文書データベース（4）に格納された文書の文書構成要素である検索対象とに従って行い、前記文書検索手段（5）は、検索の際、各検索キーワードについて文書データベース（4）の文書の該当検索対象のみを検索することを特徴とする請求項1記載の文書検索装置。
【発明の詳細な説明】
【0001】
【産業上の利用分野】本発明は、文書を蓄積した文書データベースから、利用者により入力された文書と類似の内容を持つ文書を検索するための文書検索装置に関し、特に、定型的な構造を持つ入力文書と類似の内容を持つ文書を検索するための文書検索装置に関する。
【0002】
【従来の技術】近年、文書資源のデータベース化の進展に伴って、蓄積された文書情報を効率的に再利用するための手段が要求されている。例えば、ＱＡ（質問応答）サービス業務においては、過去のＱＡ事例をデータベース化しておき、新たに受けた質問に対して、その質問と類似の質問を持つＱＡ事例をデータベースの中から簡単に見つけることができるならば、業務の大幅な効率化が期待できる。
【0003】通常、ＱＡサービス業務では、顧客からの質問自体も受付窓口で一定の型式に文書化される。したがって、このような業務に、文書データベースシステムを導入した場合、与えられた文書と類似した内容の文書を探すといった目的で利用されることになるため、文書そのものを検索キーとして類似文書を探す文書検索装置が必要である。
【0004】従来の文書検索装置においては、単語単位の検索キーと各検索キーによる検索結果間の集合演算方法とを、検索式として与えることにより検索を行っていた。例えば、“文書”と“検索”という二つの単語を両方とも含む文書を検索する場合には、“文書”ＡＮＤ

2

“検索”というような検索式を、利用者自身が入力しなければならない。
【0005】また、一つの検索式に対して複数の検索結果がある場合、全ての検索結果は同等に出力され、各検索結果の優劣を判断するための情報は出力されない。
【0006】
【発明が解決しようとする課題】以上説明したような従来の文書検索装置を、与えられた文書と類似の文書を探すという目的で利用する場合には、あらかじめ利用者自身が、その文書を特徴づける単語を検索キーとして用意する必要がある。しかし、与えられた文書と類似の文書を漏れなく探すためには、様々な観点からの単語を用意しなければならず、検索キーの数は非常に多くなるのが普通である。
【0007】また、類似の文書という曖昧な選択基準を表現するための検索式は、集合積や集合和などの単純な集合演算のみで表現しようとする限り、非常に複雑なものになる。簡単な例として、Ａ，Ｂ，Ｃの三つの単語を検索キーとして、この中の二つ以上の単語を含む文書を探すという条件は、集合積ＡＮＤおよび集合和ＯＲのみを使うと、次のような検索式になる。
【0008】（Ａ　ＡＮＤ　Ｂ）ＯＲ（Ａ　ＡＮＤ　Ｃ）ＯＲ（Ｂ　ＡＮＤ　Ｃ）
検索キーとする単語の数が増えると、このような検索式は組合せ論的に長くなる。したがって、利用者は、あらかじめ用意した検索キーの中から、検索式として表現可能な程度の数の検索キーを選択して検索を行い、求める結果が得られなければ、さらに別の検索キーを選択して検索を行うという試行錯誤を繰り返すことになり、必要十分な検索結果を得るのに時間がかかるという問題があった。
【0009】さらに、同じ検索キーで複数の文書が見つかった場合、その検索キーが文書中のどこに出現するかによって、類似性を判断する際の重要度が異なる。例えば、“文書検索”という単語で検索して、この単語が、章見出しの部分に含まれている文書と、本文中に含まれている文書とでは、明らかに章見出しに含まれている文書の方が、利用者にとって有用な情報である可能性が高い。
【0010】従来の文書検索装置を利用して、上記のような検索結果の優劣を判断するには、検索対象を章見出しまたは本文といった特定の文書構成要素に限定して数回に渡る検索を行うか、あるいは文書全体を対象とした検索の結果得られた文書に全て目を通す必要がある。したがって、検索結果の取捨選択に時間がかかるばかりでなく、利用者に十分な文書読解力を要求しなければならないという問題があった。
【0011】本発明は、上記問題点に鑑みなされたものであり、文書データベースから、文書そのものを検索キーとして類似文書を検索し、一回の検索で必要十分な検

索結果を得る文書検索装置を提供することを目的とする。

【0012】

【課題を解決するための手段】図1および図2の両者により本発明の原理説明図を示す。図において、1は適当なマーク付け言語を用いた入力構造化文書であり、利用者が検索キーとして入力したものである。2は検索キーワード集合生成手段であり、入力構造化文書1を解析して、類似文書検索を行う上で必要な文書構成要素のみを抽出した上で、それらの文書構成要素の内容に対して、必要に応じて自動キーワード抽出や関連語展開などを行うといった、文書構成要素の種類によって異なる規則を適用して検索キーワード集合3を生成する。

【0013】3は検索キーワード集合生成手段2によって生成された検索キーワード集合であるが、単なる検索キーワードの羅列ではなく、後述の文書検索手段5での類似文書検索が可能となるように構造化されて検索キーワードが格納されている。すなわち、入力構造化文書1にもともと含まれていた単語である主キーワード3aに、その単語を関連語などに展開して作られた展開キーワード3bがリンクされており、主キーワード3a同士も互いにリンクされている。

【0014】各検索キーワードには、その検索キーワードを生成するもととなった文書構成要素の種類などに応じて算出された、類似文書検索におけるその検索キーワードの重要性を示す重み3cが付加されている。重み3cは0から100までの間の数値であるが、一つの主キーワード系列、すなわち主キーワード3aとその展開キーワード3bの重みの中では、主キーワードの重みが最も高く、全ての主キーワードの重みの合計は100になるように調整されている。

【0015】なお、後述のデータベース4が構造化文書データベースとして構成された場合には、各主キーワード3aには、その主キーワード系列による検索の対象とすべき、構造化文書データベース4中の文書の文書構成要素名が、検索対象名3dとして格納されると良い。4は文書データベースである。なお、この文書データベースは、入力構造化文書1に使用したのと同じマーク付け言語を用いて構造化された文書が格納されるようにしても良い。

【0016】5は文書検索手段であり、検索キーワード集合3を用いて文書データベース4を検索し、その結果得られた検索結果候補6の文書と入力構造化文書1との類似性を評価するための確信度6aを算出する。すなわち、まず、検索キーワード集合3中の一つの主キーワード系列で検索を行い、その結果得られた文書は、中間検索結果5aとして一時的に格納される。この際、中間検索結果5a中の各文書の重み5bには、その文書がヒットした検索キーワードの重み3cを格納するが、一つの文書が複数の検索キーワードでヒットした場合には、それらの検索キーワードの重みの中で最も大きな値を格納する。

【0017】一つの主キーワード系列により検索が終了したら、その主キーワード系列の中間検索結果5aを現在までの検索結果候補6と比較し、現在までの検索結果候補6中に存在しない中間検索結果5a中の文書については、その文書を検索結果候補6に追加し、その文書の重み5bをそのまま確信度6aとして格納する。中間検索結果5a中の文書が現在までの検索結果候補6中に既に存在する場合は、検索結果候補6中のその文書の確信度6aに現在の検索で得た重み5bを加算する。

【0018】一つの主キーワード系列による中間検索結果5aを検索結果候補6に追加し終わったら、次の主キーワード系列について同様の検索処理を実行する。全ての主キーワード系列についての処理が終了した時点で、文書検索手段5の処理を完了する。8は検索結果選別手段であり、検索結果候補6の中から、確信度閾値7に設定された値以上の確信度6aを持つものを選択し、最終的な検索結果9として確信度9aと共に出力する。

【0019】

【作用】本発明における入力構造化文書1は、ISO8879で制定されたSGML(Standard Generalized Markup Language)などのマーク付け言語を利用して構造化したものである。すなわち、文書の表題、章題、本文といった文書構成要素の名前とその範囲が、適当な記号を用いて文書中にマーク付けされている。このような構造化の採用により、文書構造を考慮した検索が容易に実現可能となる。

【0020】検索キーワード集合生成手段2では、入力構造化文書1の文書構成要素の種類に応じて、その検索キーワードに重要性に応じた重み3cが付加されるといった一連の処理により、類似文書検出のための検索キーワード集合3が自動的に生成される。したがって、利用者は、どのような検索キーワードを用いてどのような手順で検出すべきかといった問題を意識することなく、文書そのものを検索キーとして入力するだけで、類似文書の検索を行うことができる。

【0021】文書検索手段5により出力される検索結果候補6の確信度6aは、検索キーワード集合3の構造と文書検索手段5の処理方法によって、0から100までの間の数値となり、確信度6aが大きい文書ほど入力構造化文書1との類似性が高いと判断することができる。例えば、もし入力構造化文書1から直接抽出された全ての主キーワード3aがその文書に含まれているなら、全ての主キーワードの重みの合計は100になるように調整されているから、その文書の確信度6aは100である。一方、主キーワード3aではなく、展開キーワード3bでヒットした文書の確信度は、展開キーワード3bの重みが主キーワード3aの重み以下に設定されているから、その分だけ確信度6aは小さくなる。

5

【0022】確信度6aは以上のようにして得られるのであるから、確信度6aが小さいほど、その文書の内容は入力構造化文書1の内容と相違していると考えることができる。確信度6aの非常に小さい文書は利用者が必要としない文書である可能性が高い。一般的には、検索結果候補6の大部分が確信度の小さい文書であるので、全ての検索結果候補6をそのまま検索結果候補9として出力することは利用者にとって好ましくない。

【0023】そこで、検索結果選別手段8では、検索結果6の中から、適当な方法で決められた確信度閾値7に設定された値以上の確信度6aを持つ文書を選別し、これを最終的な検索結果9として出力する。したがって、利用者にとって不必要な検索結果が大量に出力されるといった問題を避けることができ、類似文書検索の結果として必要十分な検索結果を出力することができる。

【0024】検索結果9は、確信度9aが付加されて出力されるので、利用者は確信度9aを参照することにより、検索結果の取捨選択を効率的に行うことができる。また、文書データベース4を構造化文書データベースとし、入力構造化文書1に使用したのと同じマーク付け言語を用いて構造化された文書が格納されるようにした場合には、さらに正確に類似性を判断することができる。

【0025】すなわち、検索キーワードの重み付けを、入力文書1の文書構成要素と、前記文書データベース4に格納された文書の文書構成要素である検索対象の両方に従って行う。さらに、検索キーワード集合3の各主キーワード3aに対してその主キーワード系列による検索の対象とすべき、構造化文書データベース4中の文書の文書構成要素名を検索対象名3dとして格納する。

【0026】そして、文書検索手段5は、構造化文書データベース4を検索する際、各検索キーワードと検索対象名3dを用いて検索する。これにより、関連する文書構成要素で検索キーワードが一致した文書に高い確信度9aが与えられる。

【0027】

【実施例】図3および図4の両者により、本発明を自動QA装置に適用した例の概略図を示す。図中、前記図1および図2で示したものと同一のものは同一の符号を付している。10は検索属性定義情報であり、入力構造化文書1中の各文書構成要素から検索キーワード集合3を生成する際に、どのような規則を適用するかなどを文書構成要素の種類ごとに定義したものであり、外部より変更可能なものである。

【0028】検索属性定義情報10は、文書構成要素名10aと適用規則名10bと検索対象名10cと相対重み10dとから構成される。文書構成要素名10aは、検索キーワード集合3を生成するもととなる入力構造化文書1中の文書構成要素名である。適用規則名10bは、文書構成要素名10aで指定される文書構成要素から検索キーワード集合3を生成する際に適用される規則名であり、検索キーワード生成規則格納手段11に格納されている規則の名前に対応し、必要に応じて複数の規則名を指定することができる。

6

【0029】検索対象名10cは、文書構成要素10aで指定される文書構成要素から生成された検索キーワードによる検索の対象とする、構造化文書データベース4中の文書の文書構成要素名であり、一つの文書構成要素名10aに対して複数の検索対象名10cを指定することができる。相対重み10dは、一組の文書構成要素名10aと検索対象名10cに対して一つ定義されるものであり、生成された検索キーワードの重要度を相対的な数値で指定する。

【0030】11は検索キーワード生成規則格納手段であり、適用規則名10bで指定される、自動キーワード抽出または関連語展開といった検索キーワード生成規則の実体が、ハードウエア、またはソフトウェアにより部品化されて格納されている。図5は、本実施例の入力構造化文書1の一例であり、顧客からの質問をISO8879の規約に従いSGML文書化したものである。各文書構成要素は“<>”で囲まれたタグによってマーク付けされている。

【0031】図6は、本実施例の構造化文書データベース4に蓄積されている文書4nの例であり、過去になされた質問に対して回答を付加したQA事例をSGML文書化したものである。本実施例は、図5のような型式の顧客からの質問文書1をそのまま検索キーとして、図4のような過去のQA事例の文書4nを蓄積したデータベースを検索し、質問に対する回答の参考になるようなQA事例を出力するものである。

【0032】以下に、図3および図4に基づき、本実施例の動作を説明する。まず、検索属性定義情報10の内容について説明する。検索属性定義情報10では、入力構造化文書1中の“表題”、“製品名”、“質問文”の三つの文書構成要素に対する検索属性が定義されている。この三つ以外の文書構成要素、例えば“質問者氏名”など類似検索を行う上で不要の情報は、検索属性定義情報10の中に含まない。

【0033】図3の例では、適用規則名10bとして、“自動キーワード抽出”、“関連語展開”の二種類が指定されている。“自動キーワード抽出”は、文章中に含まれる単語を自動的に抽出して主キーワード3aとするものであり、“表題”や“質問文”のように、自然文で記入される文書構成要素に適用される。もし一つの文書構成要素の内容から複数の単語が抽出された場合には、その個数分の主キーワード3aが生成される。

【0034】しかし、“製品名”のようにもともと決められた単語が記入される文書構成要素に対しては、“自動キーワード抽出”は適用せず、記入されている内容をそのまま主キーワード3aとすればよい。“関連語展開”は、文書構成要素の内容から直接抽出された単語を

主キーワード３ａとして、さらにその単語の関連語も展開キーワード３ｂとするものであり、類似文書検索をする上で必要な検索範囲の拡張を行うことができる。

【００３５】“自動キーワード抽出”や“関連語展開”を行うための手段は、検索キーワード生成規則格納手段１１の部品の一部として格納されているが、これらの手段の説明は本発明の目的とするところではないので省略する。検索対象名１０ｃは、本実施例の場合、基本的には、文書構成要素名１０ａと同じである。すなわち、入力構造化文書１中のある文書構成要素から生成された検索キーワードは、構造化文書データベース４中の文書の同じ文書構成要素を検索対象とする。

【００３６】しかし、入力構造化文書１中の“質問文”から生成された検索キーワードは、構造化文書データベース４中のＱＡ事例において、“回答文”の中に含まれていても関連事例である可能性があるので、“質問文”の検索対象名には、“回答文”も指定しておく。相対重み１０ｄは、質問を特徴付けるのに最も重要な文書構成要素である“表題”の相対重みを最も大きくする。“質問文”の相対重みに関しては、“回答文”を検索対象とする場合の重みを“質問文”を検索対象とする場合よりも小さく設定しておくことにより、検索対象の違いによる検索キーワードの重要性の違いを反映することができる。

【００３７】検索キーワード集合生成手段２では、以上説明した検索属性定義情報１０を参照して、検索キーワード生成規則格納手段１１に格納された規則を適用し、入力構造化文書１から検索キーワード集合３を生成する。次に、図７のフローチャートに基づいて、検索キーワード集合生成手段２での動作を説明する。

【００３８】まず、ステップＳ１１で検索属性定義情報１０の文書構成要素名１０ａを一つ読み込みステップＳ１３へ進むが、ここで読み込むべき文書構成要素名１０ａがなくなったら、ステップＳ１２からステップＳ１５へ進む。ステップＳ１３では、ステップＳ１１で読み込んだ文書構成要素名１０ａに対応する文書構成要素の内容を入力構造化文書１中から抽出する。

【００３９】ステップＳ１４では、その文書構成要素の適用規則名１０ｂに対応する検索キーワード生成規則を検索キーワード生成規則格納手段１１から呼び出し、呼び出した規則をその文書構成要素の内容に適用して、検索キーワード集合を構築していく。この際、その文書構成要素に対して複数の検索対象名１０ｃが指定されている場合には、検索対象名１０ｃのみが異なる同じ内容の主キーワード系列を、検索対象名１０ｃの個数分だけ生成する。主キーワード３ａの重み３ｃには、相対重み１０ｄを、その文書構成要素から生成された主キーワード３ａの個数で等分した値を格納する。

【００４０】展開キーワード３ｂの重み３ｃは、その系列の主キーワード３ａの重み３ｃから算出するが、適用される検索キーワード生成規則により算出方法が異なる。例えば、“関連語展開”の場合、主キーワード３ａと展開キーワード３ｂの意味関係が遠いほど、展開キーワードの重み３ｃを小さくする。ステップＳ１４での処理が終了したら、ステップＳ１１へ戻る。

【００４１】ステップＳ１５では、各検索キーワードに付加された重み３ｃの再規格化を行う。すなわち、主キーワード３ａに付加された重みの合計が１００になるような一定の定数を、全ての検索キーワードの重み３ｃに乗じる。次に、図４に戻ると、文書検索手段５では、上記手順に従って生成された検索キーワード集合３に基づき、構造化文書データベース４を検索する。

【００４２】構造化文書データベース４は、インバーテッドファイルなどの手法により、検索対象名と検索キーワードから目的の文書を検索することのできる構造とする。次に、図８、図９、図１０の３図で示すフローチャートに基づいて、文書検索手段５での動作を説明する。まず、ステップＳ２１では、検索キーワード集合３から主キーワード系列を一つ取り出し、次いでステップＳ２３へ進むが、ここで取り出す主キーワード系列がなくなったら、ステップＳ２２のＹＥＳから終了へ進み文書検索手段５での処理を終了する。

【００４３】ステップＳ２３では、ステップＳ２１で取り出した主キーワード系列の主キーワード３ａから検索対象名３ｄを取り出しておく。ステップＳ２４では、ステップＳ２２で取り出した主キーワード系列中の検索キーワード集合をリンクされた順序に従って一つ取り出しステップＳ２６へ進むが、ここで取り出す検索キーワードがなくなったら、ステップＳ２５からステップＳ３３へ進む。

【００４４】ステップＳ２６では、ステップＳ２３で取り出した検索対象名３ｄと、ステップＳ２４で取り出した検索キーワードで、構造化文書データベース４を検索する。ステップＳ２７では、ステップＳ２６で検索した結果から、一つの構造化文書を取り出し、ステップＳ２９へ進むが、ここで取り出す文書がなくなったら、ステップＳ２８からステップＳ２４へ戻る。

【００４５】ステップＳ２９では、ステップＳ２７で取り出した構造化文書が既に中間検索結果５ａ中に存在する文書かどうかが判定され、存在する文書ならばステップＳ３１へ進み、新規な文書であればステップＳ３０へ進む。ステップＳ３０では、その構造化文書を中間検索結果５ａに追加すると共に、現在の検索キーワードの重み３ｃをその構造化文書の重み５ｂに格納して、ステップＳ２７へ戻る。

【００４６】ステップＳ３１では、中間検索結果５ａ中の現在の検索結果と同一の文書の重み５ｂと、現在の検索キーワードの重み３ｃを比較し、現在の検索キーワードの重み３ｃの方が大きければステップＳ３２へ進み、そうでなければステップＳ２７へ戻る。ステップＳ３２

では、中間検索結果５ａ中の現在の検索結果と同一の文書の重み５ｂを現在の検索キーワードの重み３ｃに置き換えて、ステップＳ２７へ戻る。

【００４７】ステップＳ３３では、中間検索結果５ａ中の文書を一つ取り出しステップＳ３５へ進むが、ここで取り出す文書が無くなったら、ステップＳ３４からステップＳ３８へ進む。ステップＳ３５では、ステップＳ３３で取り出した構造化文書が既に検索結果候補６中に存在するかどうかを調べ、新規の文書であればステップＳ３６へ進み、既に検索結果候補６中に存在する文書ならばステップＳ３７へ進む。

【００４８】ステップＳ３６では、その構造化文書を検索結果候補６に追加すると共に、中間検索結果５ａでの重み５ｂをその構造化文書の確信度６ａに格納して、ステップＳ３３へ戻る。ステップＳ３７では、中間検索結果５ａ中でのその文書の重み５ｂを、検索結果候補６中でのその文書の確信度６ａに加算し、ステップＳ３３へ戻る。

【００４９】ステップＳ３８では、中間検索結果５ａの内容を消去し、ステップＳ２１へ戻る。再び図４に戻ると、上記文書検索手段５の処理手順によって、検索結果候補６が作成されるが、確信度６ａの非常に小さい文書は、入力した質問と無関係の内容である可能性が高いので、そのような文書を検索結果選別手段８で削除する。

【００５０】すなわち、検索結果選別手段８では、検索結果６の中から、適当な方法で決められた確信度閾値７に設定された値以上の確信度６ａを持つ文書を選別し、これを最終的な検索結果９として確信度９ａと共に出力する。このように、本実施例の自動ＱＡ装置は、質問文書をそのまま入力するだけで、その質問に対する回答を得る上で参考になる必要十分な量のＱＡ事例を検索結果として得ることができるものである。

【００５１】なお、本発明の文書検索装置は、上記実施例のようなＱＡ事例の検索に対してのみではなく、例えば特許文書などの定型的な文書構造を持つ文書の類似検索全てに対して適用可能である。また、上記実施例では、検索キーワードを生成する際の適用規則として、“自動キーワード抽出”および、“関連語展開”のみを使用していたが、必要に応じて、半角と全角を全角に統一するといったキーワード表記の正規化など他の規則を組み込むことができる。

【００５２】さらに、本発明は、検索属性定義情報１０の検索対象名１０ｃおよび検索キーワード集合３の検索対象名３ｄを省略することが可能である。以上説明したように、定型的な構造を持つ文書を蓄積した文書データベースの類似文書検索において、利用者が検索キーワードや検索手順等を何ら意識しなくても、文書そのものを検索キーとして入力するだけで、文書構造に応じた検索キーワード集合が内部的に生成され、一回の検索で必要十分な検索結果を得ることができる。

【００５３】さらに、検索結果には、入力文書と類似性を示す確信度が付加されているため、検索結果の取捨選択を効率的に行うことができることから、類似文書検索装置の機能向上に寄与するところが大きい。

【００５４】

【発明の効果】以上説明したように、本発明の方法によれば、文書データベースから、文書そのものを検索キーとして類似文書を検索し、一回の検索で必要十分な検索結果を得ることができる。

【図面の簡単な説明】

【図１】本発明の文書検索装置の原理説明図（その１）。

【図２】本発明の文書検索装置の原理説明図（その２）。

【図３】本発明の文書検索装置の実施例を示す概略図（その１）。

【図４】本発明の文書検索装置の実施例を示す概略図（その２）。

【図５】図３の入力文書の一例を示す図。

【図６】図４のデータベースに蓄積される文書の一例を示す図。

【図７】図３の検索キーワード集合生成手段の動作を説明するフローチャート。

【図８】図４の文書検索手段の動作を説明するフローチャート（その１）。

【図９】図４の文書検索手段の動作を説明するフローチャート（その２）。

【図１０】図４の文書検索手段の動作を説明するフローチャート（その３）。

【符号の説明】

１…入力構造化文書
２…検索キーワード集合生成手段
３…検索キーワード集合
３ａ…主キーワード
３ｂ…展開キーワード
３ｃ…重み
３ｄ…検索対象名
４…文書データベース
５…文書検索手段
５ａ…中間検索結果
５ｂ…重み
６…検索結果候補
６ａ…確信度
７…確信度閾値
８…検索結果選別手段
９…検索結果
９ａ…確信度
１０…検索属性定義情報
１０ａ…文書構成要素名
１０ｂ…適用規則名

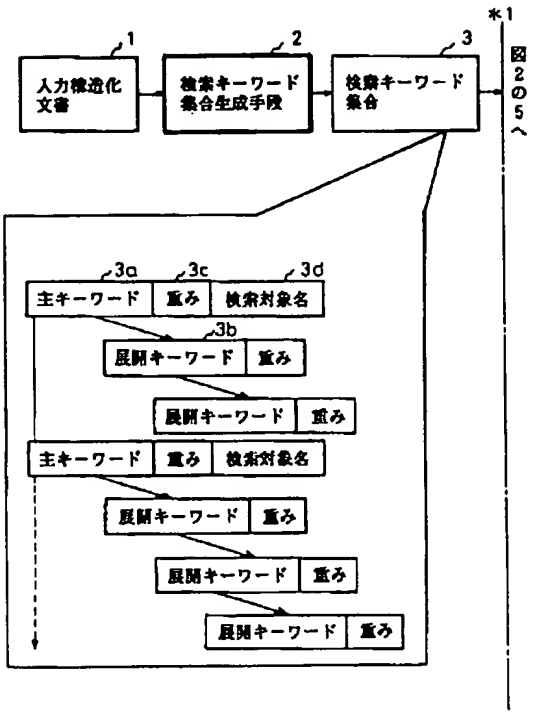
11

10c…検索対象名
10d…相対重み

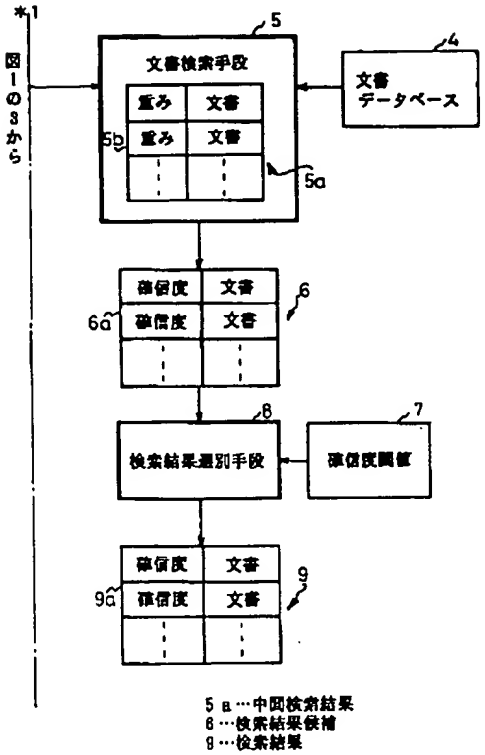
12

11…検索キーワード生成規則格納手段

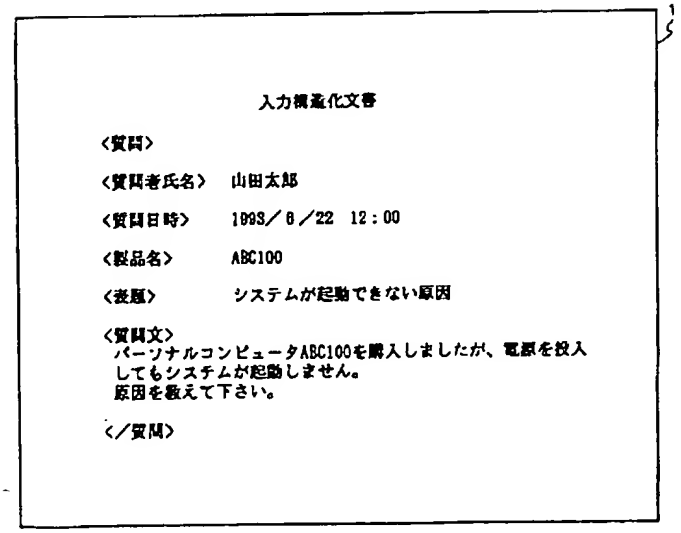
【図1】



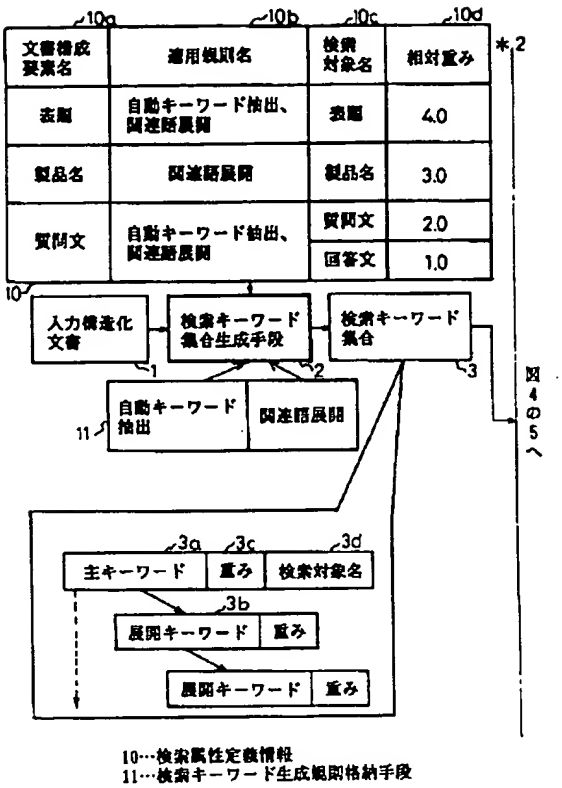
【図2】



【図5】

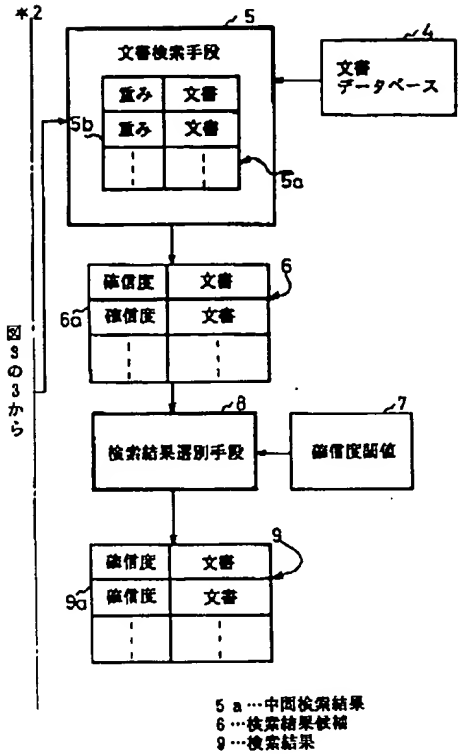


【図3】



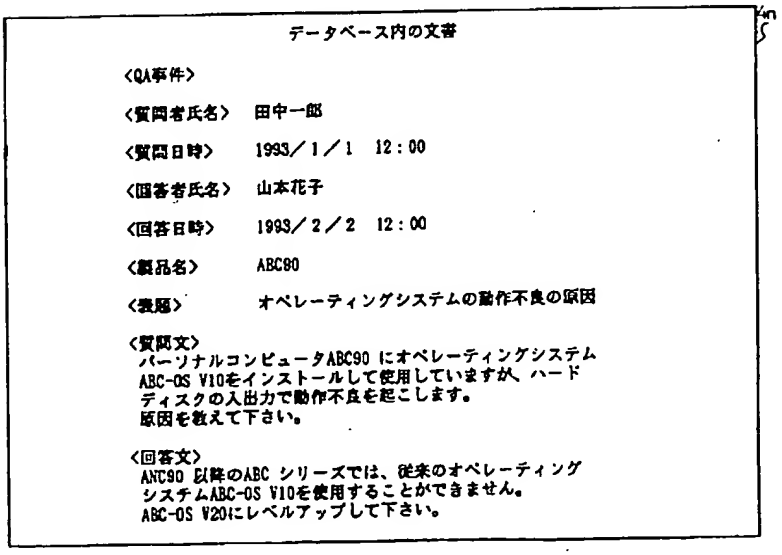
10…検索属性定義情報
11…検索キーワード生成規則格納手段

【図4】



5a…中間検索結果
6…検索結果候補
9…検索結果

【図6】



データベース内の文書

〈QA事件〉

〈質問者氏名〉　田中一郎

〈質問日時〉　1993／1／1　12：00

〈回答者氏名〉　山本花子

〈回答日時〉　1993／2／2　12：00

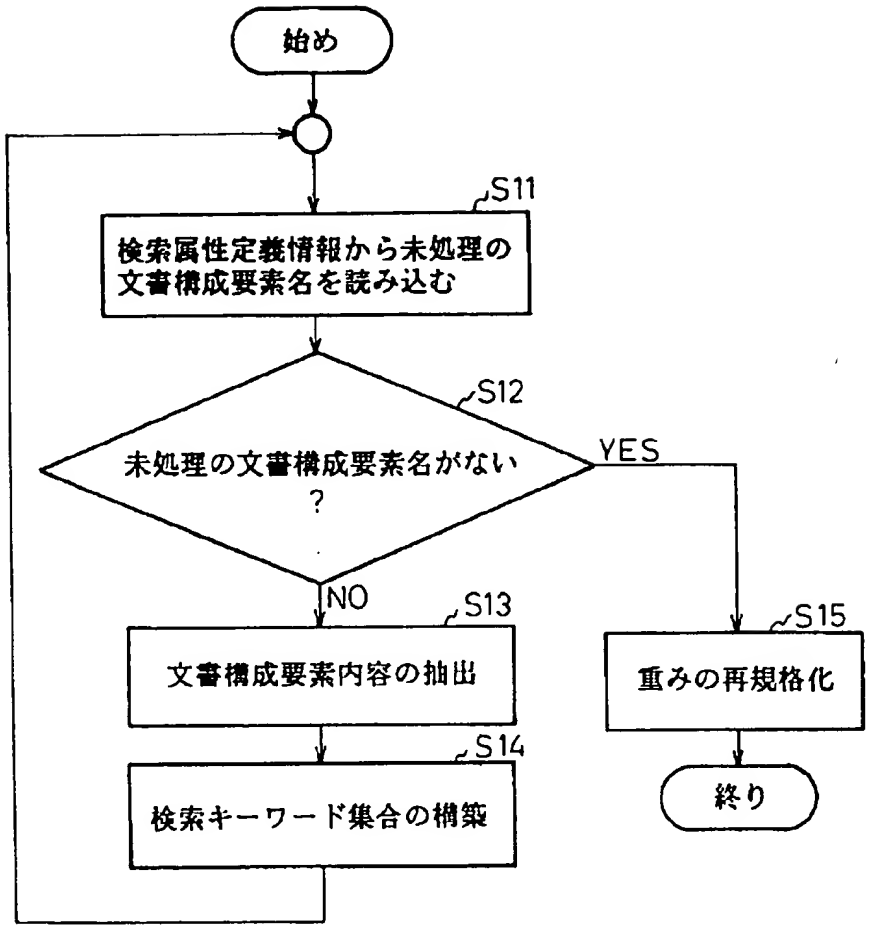
〈製品名〉　ABC90

〈表題〉　オペレーティングシステムの動作不良の原因

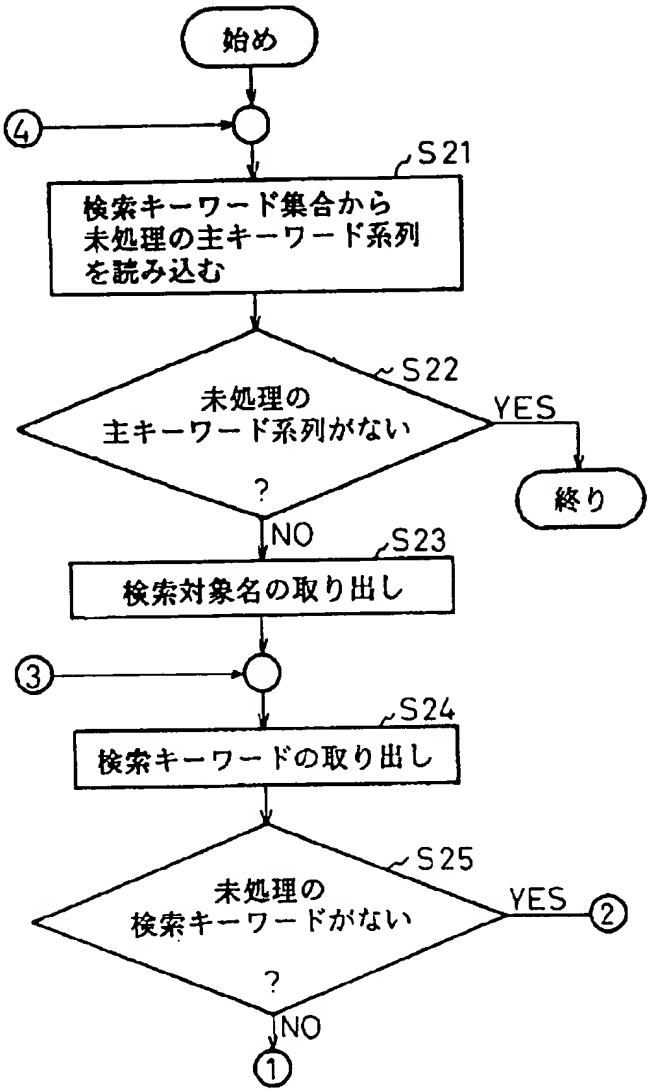
〈質問文〉
パーソナルコンピュータABC90 にオペレーティングシステム ABC-OS V10をインストールして使用していますが、ハードディスクの入出力で動作不良を起こします。
原因を教えて下さい。

〈回答文〉
ANC90 以降のABC シリーズでは、従来のオペレーティングシステムABC-OS V10を使用することができません。
ABC-OS V20にレベルアップして下さい。

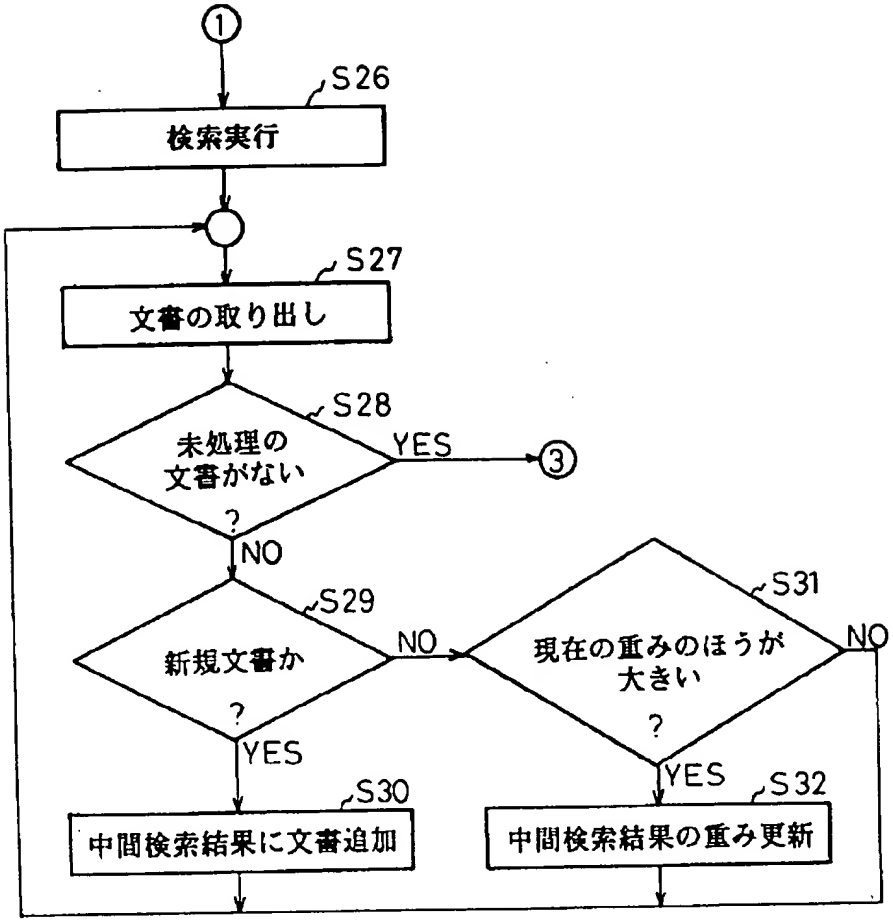
【図7】


始め
S11
検索属性定義情報から未処理の文書構成要素名を読み込む
S12
未処理の文書構成要素名がない？
YES
NO
S13
文書構成要素内容の抽出
S14
検索キーワード集合の構築
S15
重みの再規格化
終り

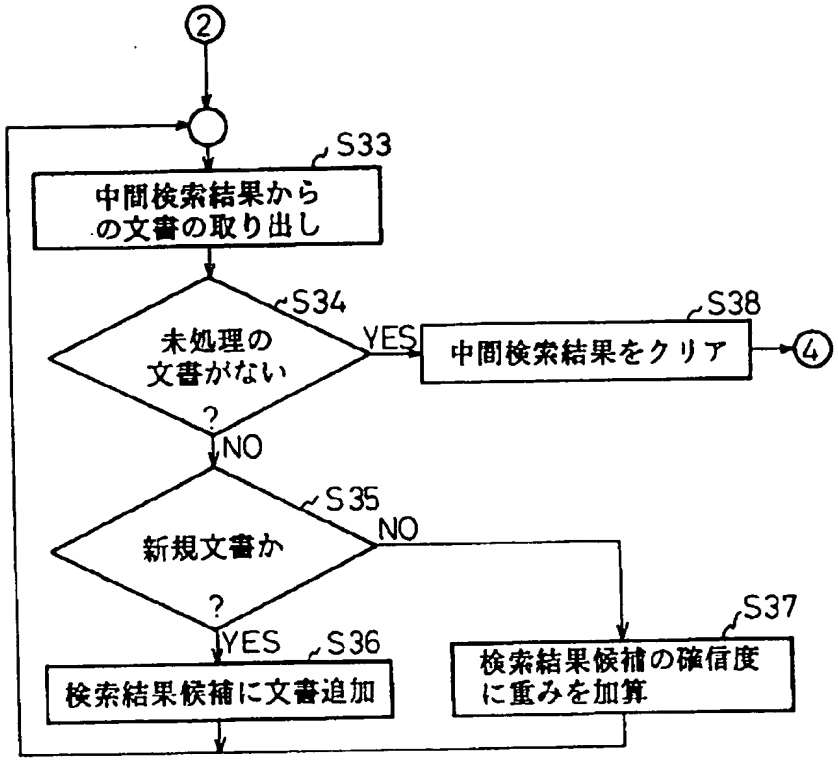
【図8】


始め
④
S21
検索キーワード集合から
未処理の主キーワード系列
を読み込む
S22
未処理の
主キーワード系列がない
?
YES
終り
NO
S23
検索対象名の取り出し
③
S24
検索キーワードの取り出し
S25
未処理の
検索キーワードがない
?
YES
②
NO
①

【図9】


①
S26
検索実行
S27
文書の取り出し
S28
未処理の文書がない
YES
③
NO
S29
新規文書か
NO
S31
現在の重みのほうが大きい
NO
YES
YES
S30
中間検索結果に文書追加
S32
中間検索結果の重み更新

【図10】


②
S33
中間検索結果から
の文書の取り出し
S34
未処理の
文書がない
?
YES
S38
中間検索結果をクリア
④
NO
S35
新規文書か
?
NO
YES
S36
検索結果候補に文書追加
S37
検索結果候補の確信度
に重みを加算